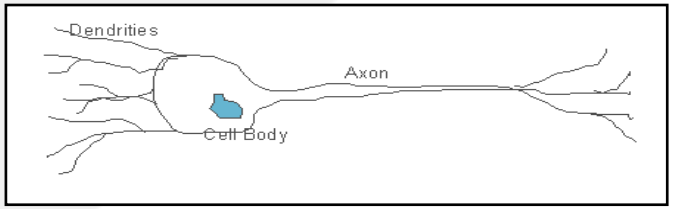
분류 및 예측 (2) :
신경망(Neural Network)

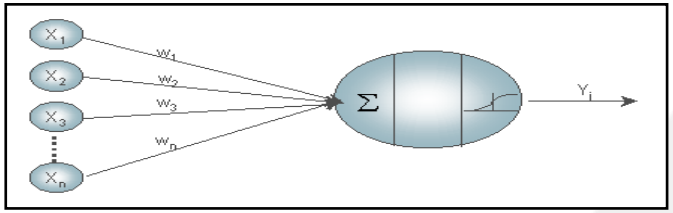
신경망(Neural Networks)의 개요

- 데이터 마이닝 알고리즘 중 가장 많이 알려진 것이 신경망 분석이며, 보통 데이터 마이닝에서 "신경망 분석 = 패턴을 찾아내는 것"이라고 연상할 만큼 잘 알려진 분석이다.

- 인간 두뇌의 신경망을 흉내 내어 실제 자신이 가진 데이터로부터의 반복적인 학습 과정을 거쳐 데이터에 숨어 있는 패턴을 찾아내는 모델링 기법



신경세포(neuron)



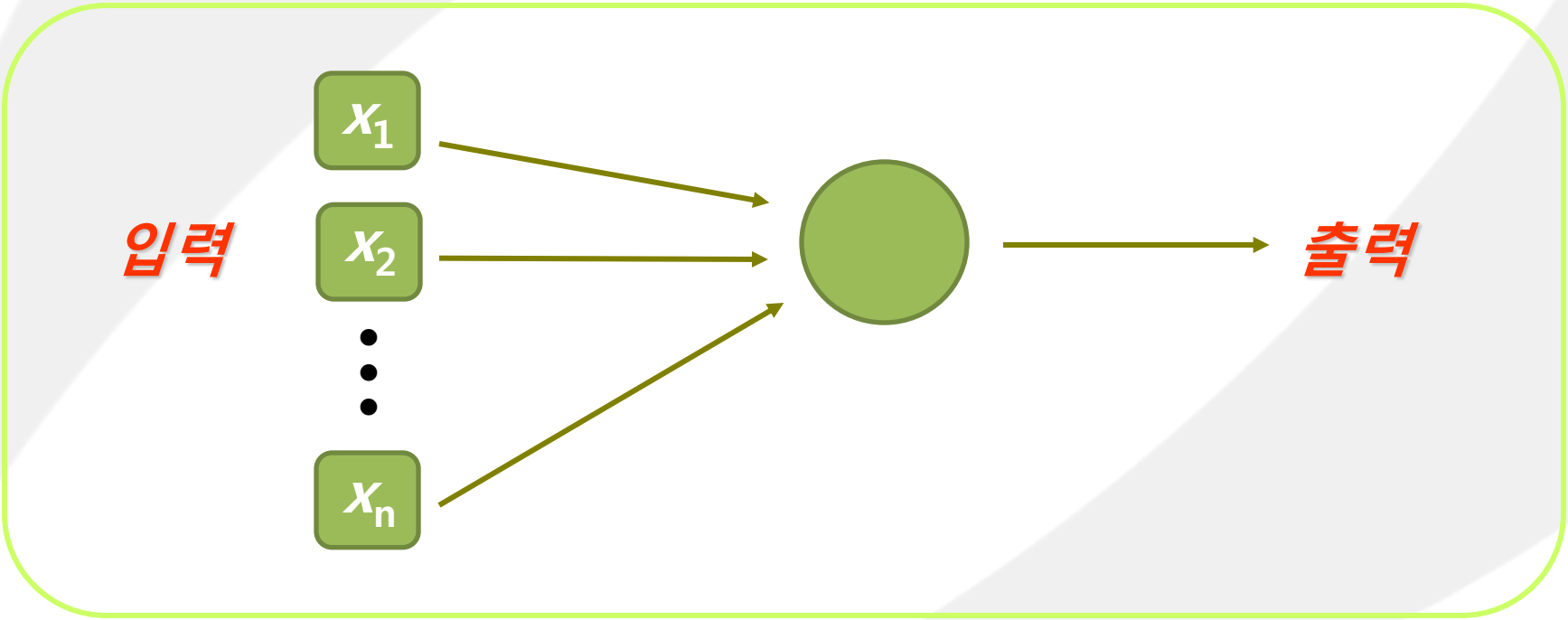
신경망(neural networks)

- 계층 구조를 갖는 수많은 프로세싱 요소로 이루어진 수학모형
  - ✓ 신경망 이론의 다양한 아키텍처를 이용하여 예측모델 생성
  - ✓ 자료의 패턴이 변화함에 따라 이를 학습하고, 이에 가중치를 변화 적용하여, 최적의 해를 구함

- 장단점
  - ✓ 비선형 자료, 범주/연속형 혼합 자료 처리가 탁월하고 통계적 가정이 불필요
  - ✓ 설명변수들이 목표변수에 구체적으로 어떠한 영향을 주는지 해석하기 어렵고, Over-Fitting 가능성 높음

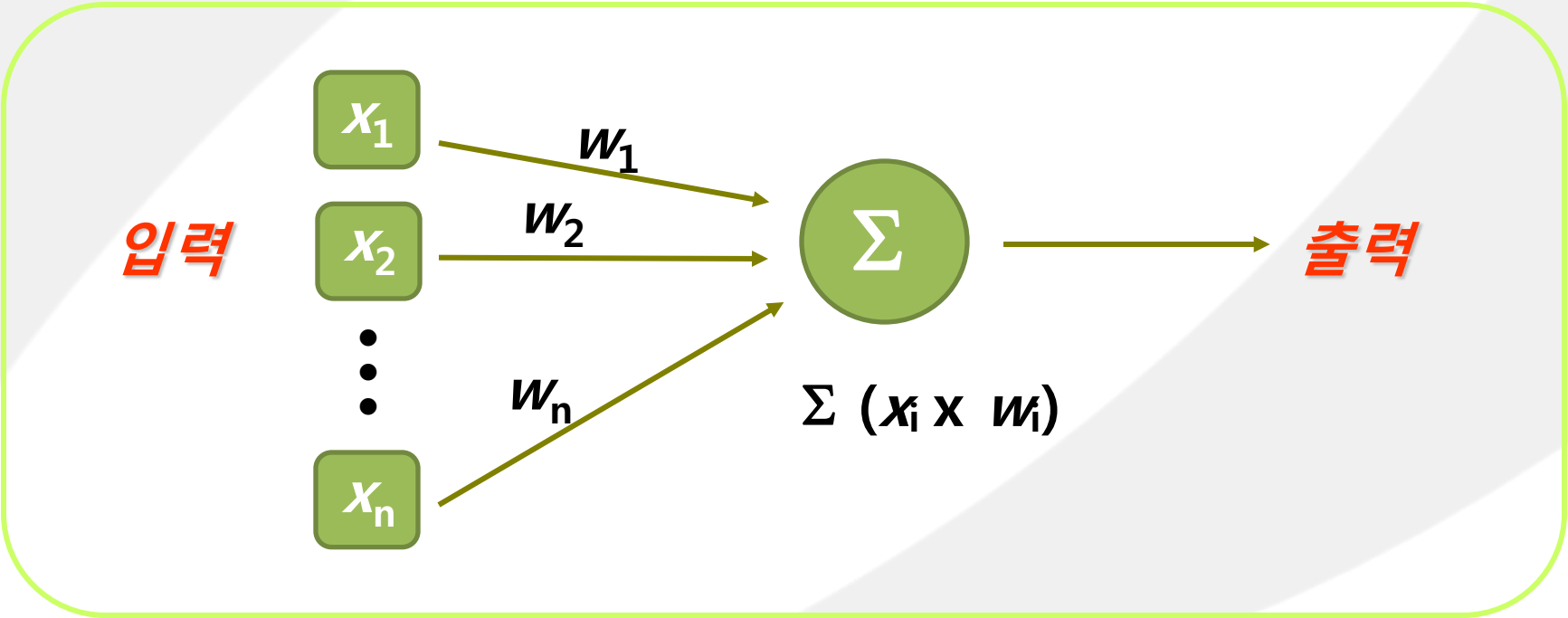
신경망(Neural Networks)의 구성요소 (1/3)

- 프로세싱 노드(processing unit, node)
  - 입력신호를 측정
  - 총 입력신호를 합산
  - 출력신호를 결정



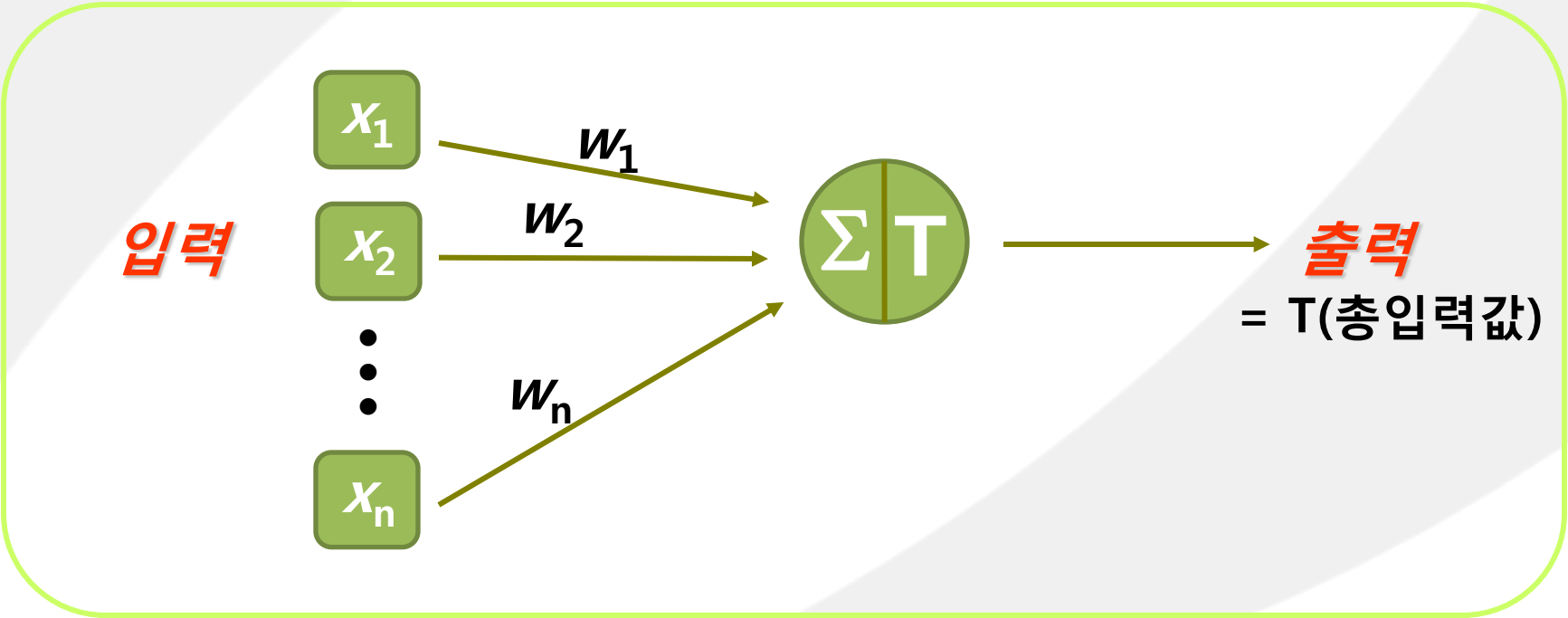
신경망(Neural Networks)의 구성요소 (2/3)

- 연결강도(weight)
  - 입력신호의 강도를 표현
- 총 입력값
  - 입력값의 선형결합함수
  - 총 입력값 = $x_1 \times w_1 + x_2 \times w_2 + \ldots + x_n \times w_n$



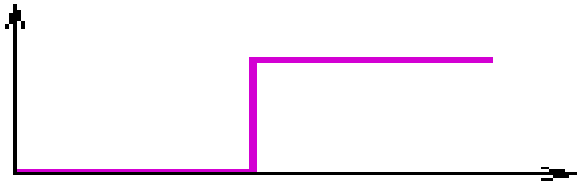
신경망(Neural Networks)의 구성요소 (3/3)

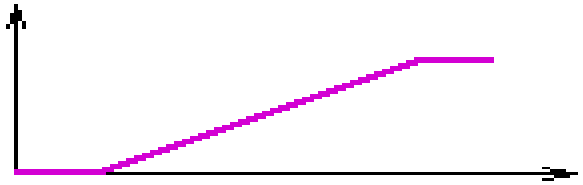
- 활성함수(activation/transfer function)
  - 입력정보의 합성값(결합값)을 일정 범위의 값으로 전환해주는 함수
  - 출력값을 결정
  - 비선형 함수를 사용

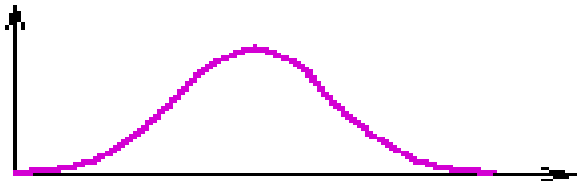


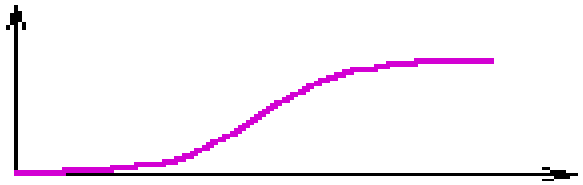
Activation Functions

- 총입력값 (a = ∑ xi * wi)가 제한된 범위를 취하도록 하는 선형 또는 비선형 함수



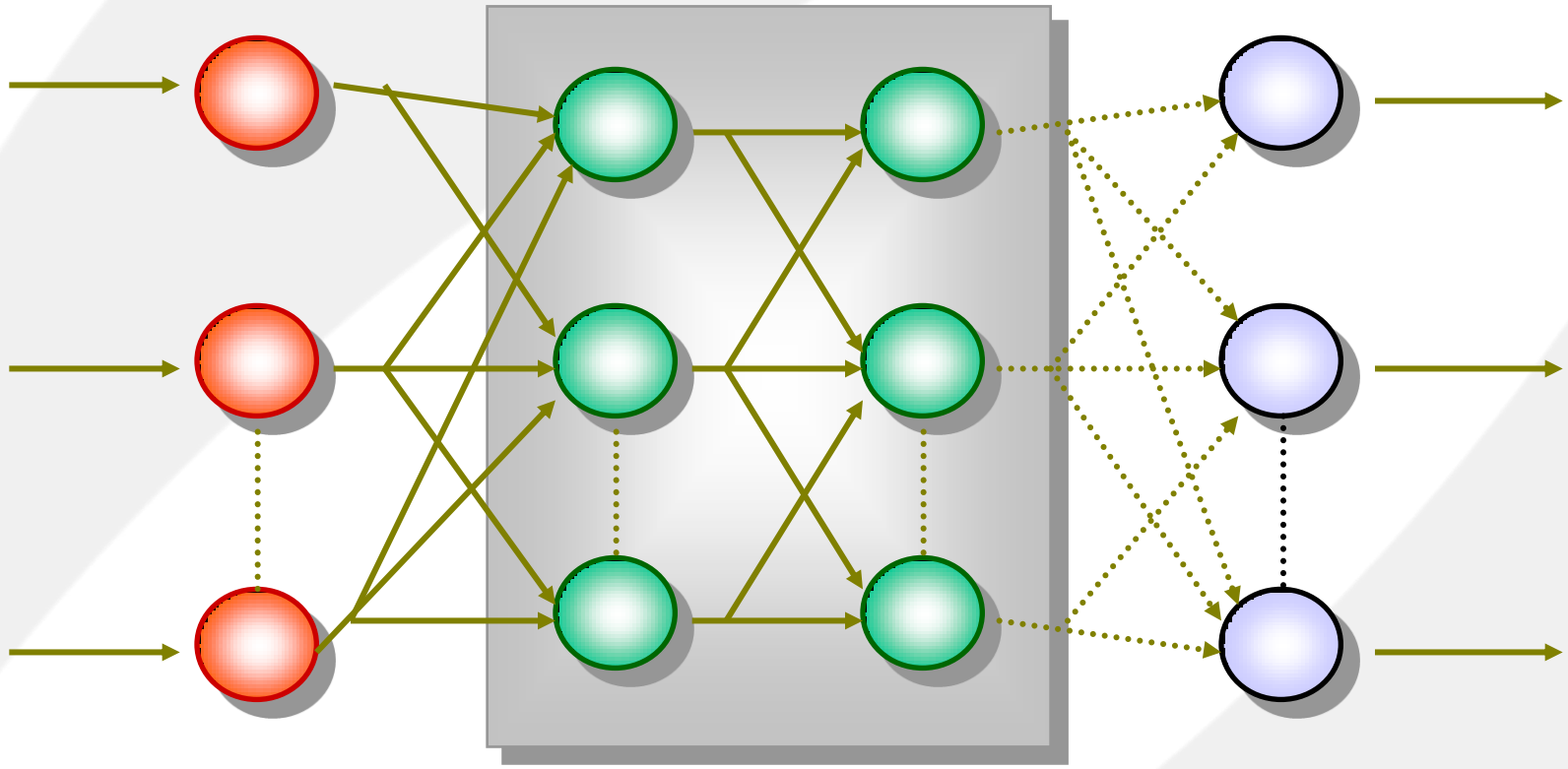






$\text{logistic}(x) = 1/(1+e^{-x})$

신경망의 계층구조



입력계층
(input layer)

은닉계층
(hidden layer)

출력계층
(output layer)

MLP (Multi-Layer Perceptron)

다층 퍼셉트론 (Multi-Layer Perceptron, MLP) 구조

- 입력층 뉴런(입력변수)로부터 전달되는 신호들을 모아 선형결합
- $X_1, \cdots, X_p$ 를 설명변수(입력 노드)라고 할 때 다음 뉴런에

$$L = w_1 X_1 + \cdots + w_p X_p$$

이 전달된다. 여기서 $w_1, \cdots, w_p$ 는 신경선(synapse)에 붙는 가중값(weight)

- 뉴런의 활성화
  - 로지스틱(logistic) : $S = e^L / (1 + e^L), \quad 0 \le S \le 1$
  - 쌍곡 탄젠트(hyperbolic tangent) : $S = (e^L - e^{-L}) / (e^L + e^{-L}), \quad -1 \le S \le 1$
- 출력노드
  - 연속형 : $O = L$
  - 범주형 : 소프트맥스(softmax)

$$O_k = \frac{\exp(L_k)}{\sum_{j=1}^{K} \exp(L_j)}$$

여기서 k는 범주.

RBF (Radial Basis Function)

방사형 기저함수(Radial Basis Function, RBF) 구조

- 입력층 뉴런(입력변수)로부터 전달되는 신호들을 모아 중심신호와의 거리에 역비례하는 강도로 변환

$$R = \exp\left[-\frac{1}{2\sigma^2}\{(X_1-\mu_1)^2+\cdots+(X_p-\mu_p)^2\}\right]$$

여기서 $\mu_1,\cdots,\mu_p$ 는 각 신호의 중심

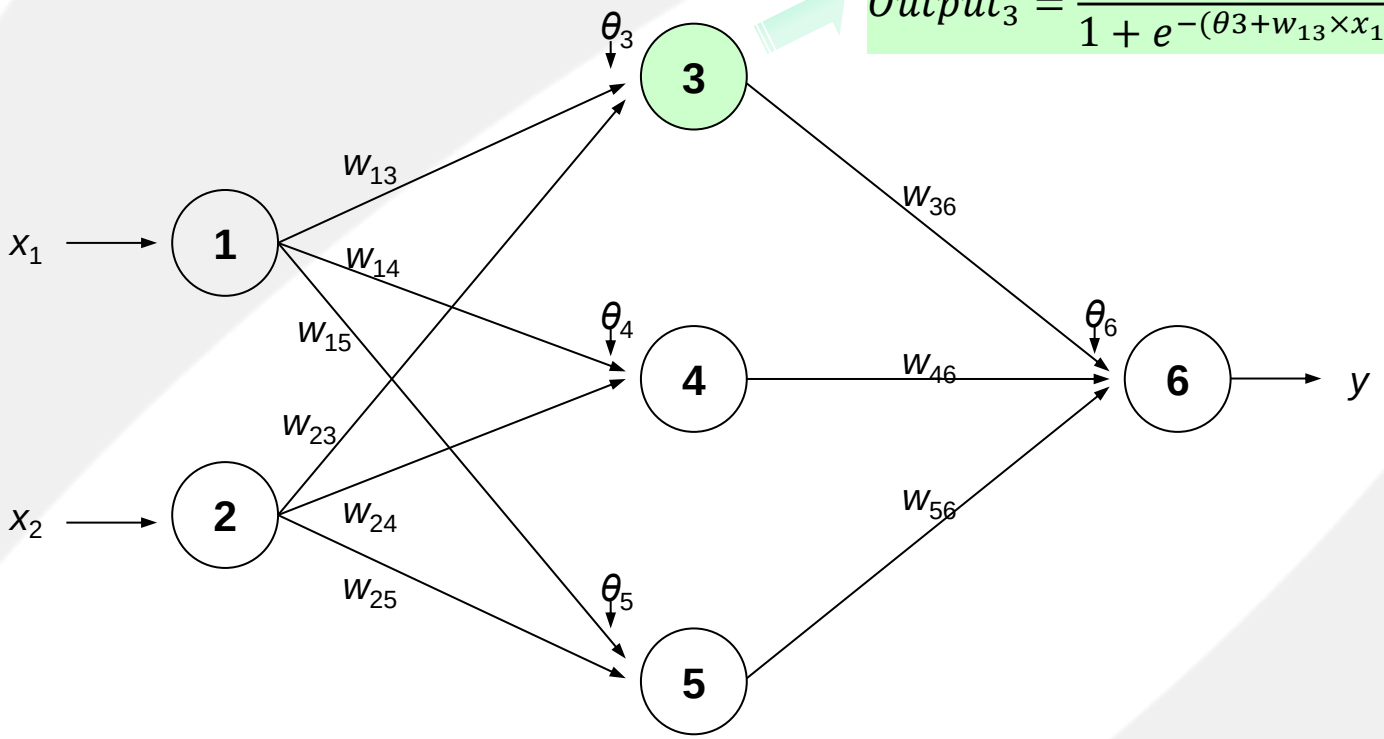
- 출력노드: $L = w_1R_1+\cdots+w_JR_J$

  - 연속형 : $O = L$
  - 범주형: 소프트맥스(softmax)

$$O_k = \frac{\exp(L_k)}{\sum_{j=1}^{K}\exp(L_j)}$$

ANN Example (1/2)

$$Output_3 = \frac{1}{1 + e^{-(\theta 3 + w_{13} \times x_1 + w_{23} \times x_2)}}$$

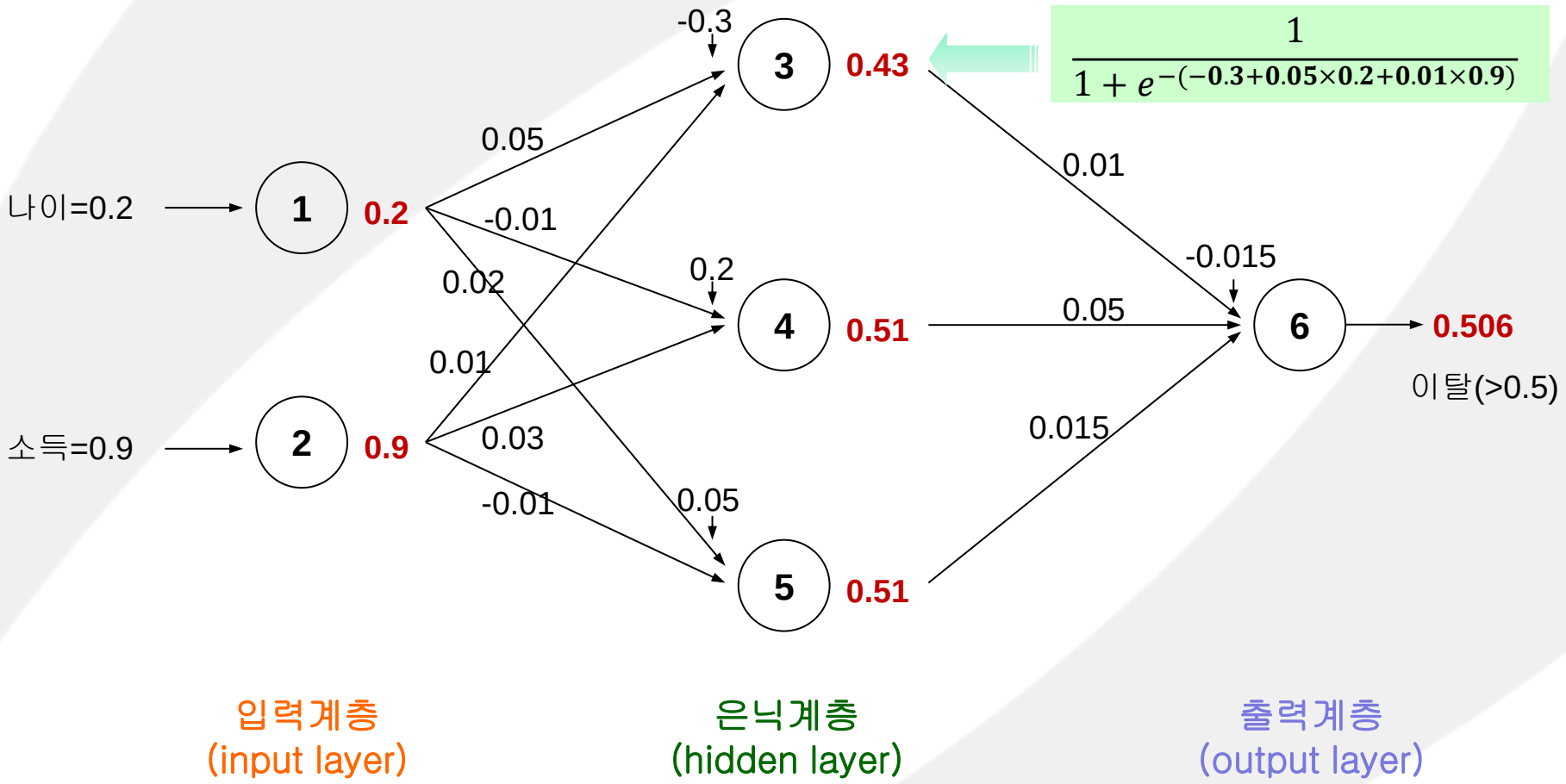


입력계층
(input layer)

은닉계층
(hidden layer)

출력계층
(output layer)

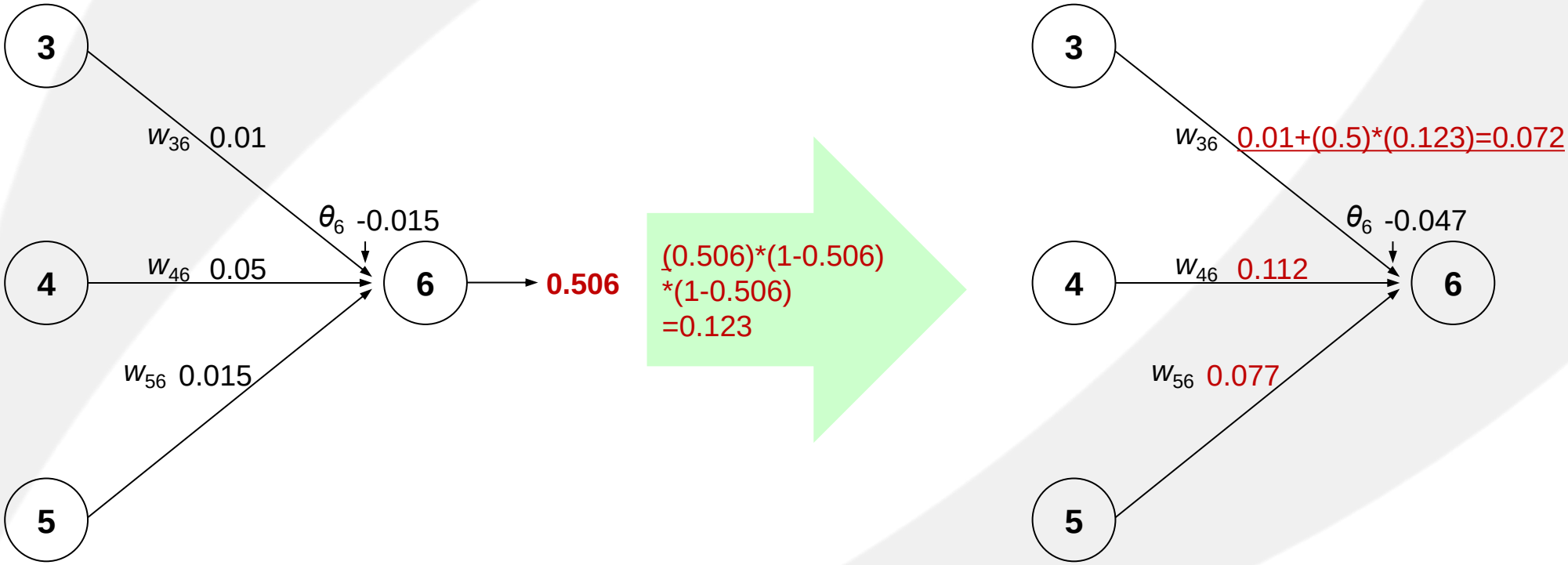
ANN Example (2/2)



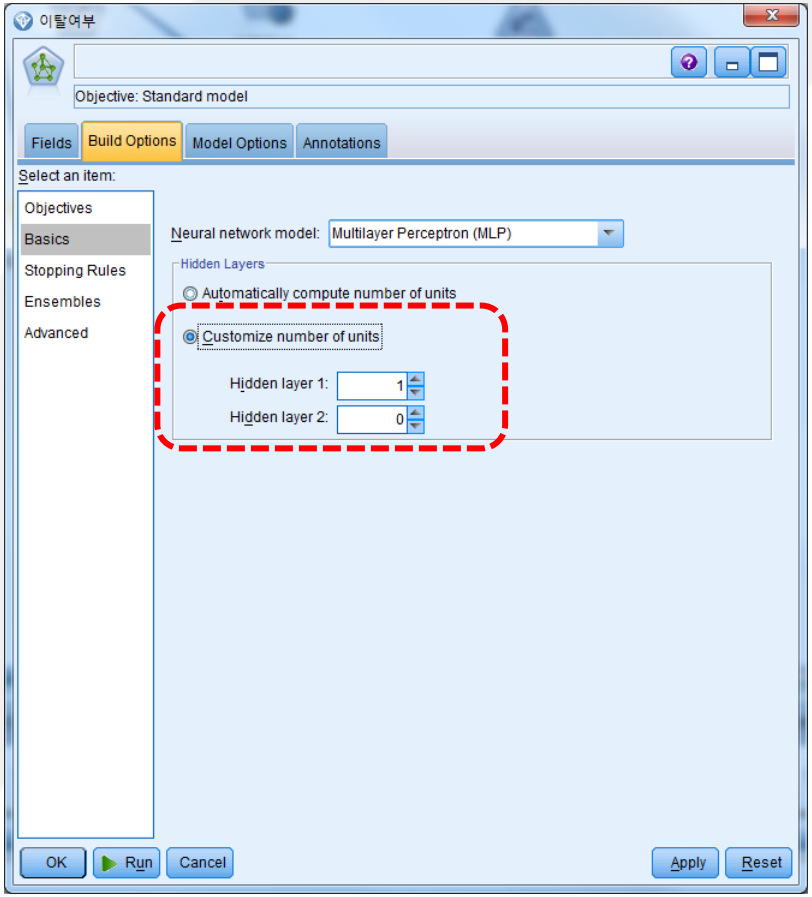
Learning method: Back propagation

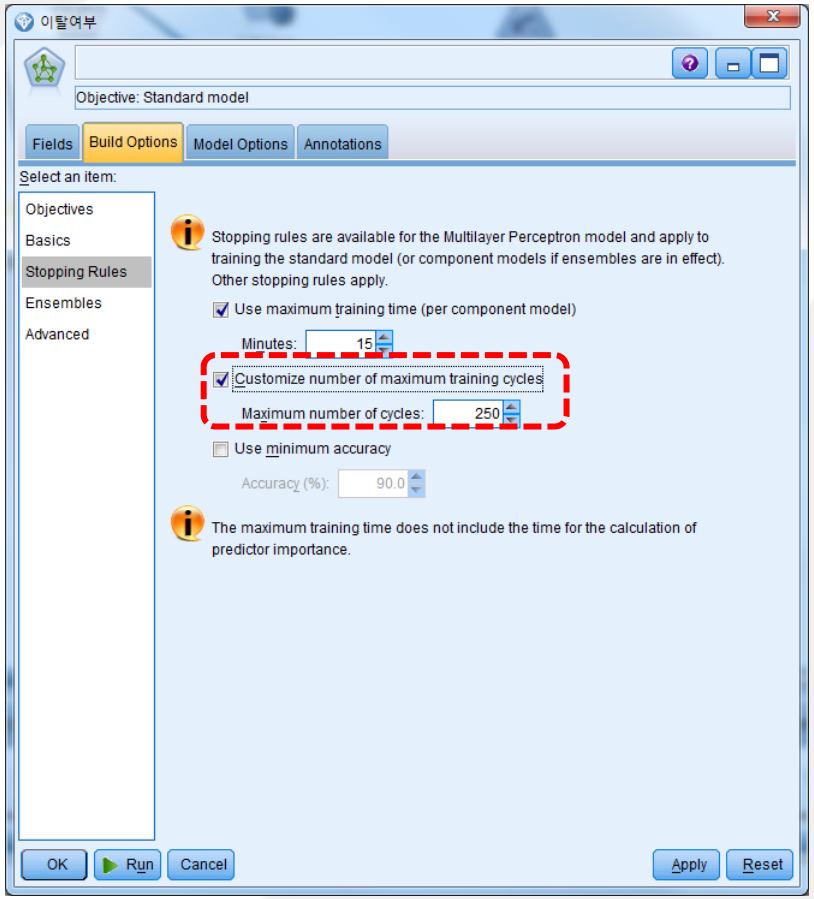
Learning NN – adjusting weights to minimize error (E)

(출력노드 오차) = (출력노드 값) * (1−출력노드 값) * (실측값− 출력노드 값)
(새로운 연결강도) = (현재 연결강도) + (학습률) * (출력노드 오차)



Neural Net 노드





- 독립변수와 종속변수 개수의 합을 n이라 할 때 n/2, n, 3n/2, 2n의 총 4가지 경우를 일반적으로 설정

- 학습용 데이터 개수의 50-100배를 일반적으로 적용

Neural Net 실습 (1/3)

신경망 분석

상황

국내 홈쇼핑 A사는 최근 소비자의 반품 횟수가 증가됨에 따라 마케팅 부서의 김팀장이 반품 고객의 특성을 파악하고자 함.

데이터

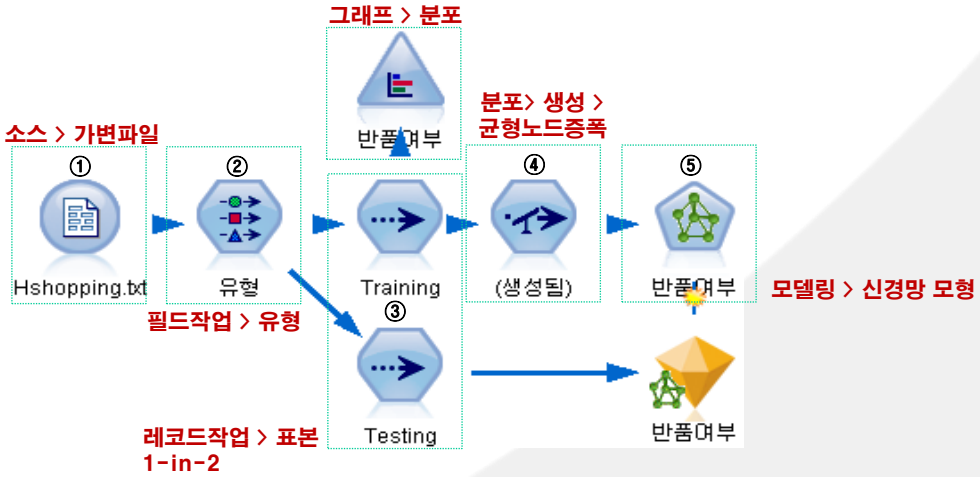
홈쇼핑 A사 고객 500명에 대한 성별, 나이, 구매금액, 홈쇼핑 출연자, 반품 여부

분석 과정

① 데이터 준비 → ② 변수 지정 → ③ 훈련 · 테스트자료 분류 → ④ 균형화 작업→ ⑤ 신경망 분석

Data: 실습 > 1. 지도학습모형 > Hshopping.txt

신경망 분석 과정



Neural Net 실습 (2/3)

▷ 실습 결과

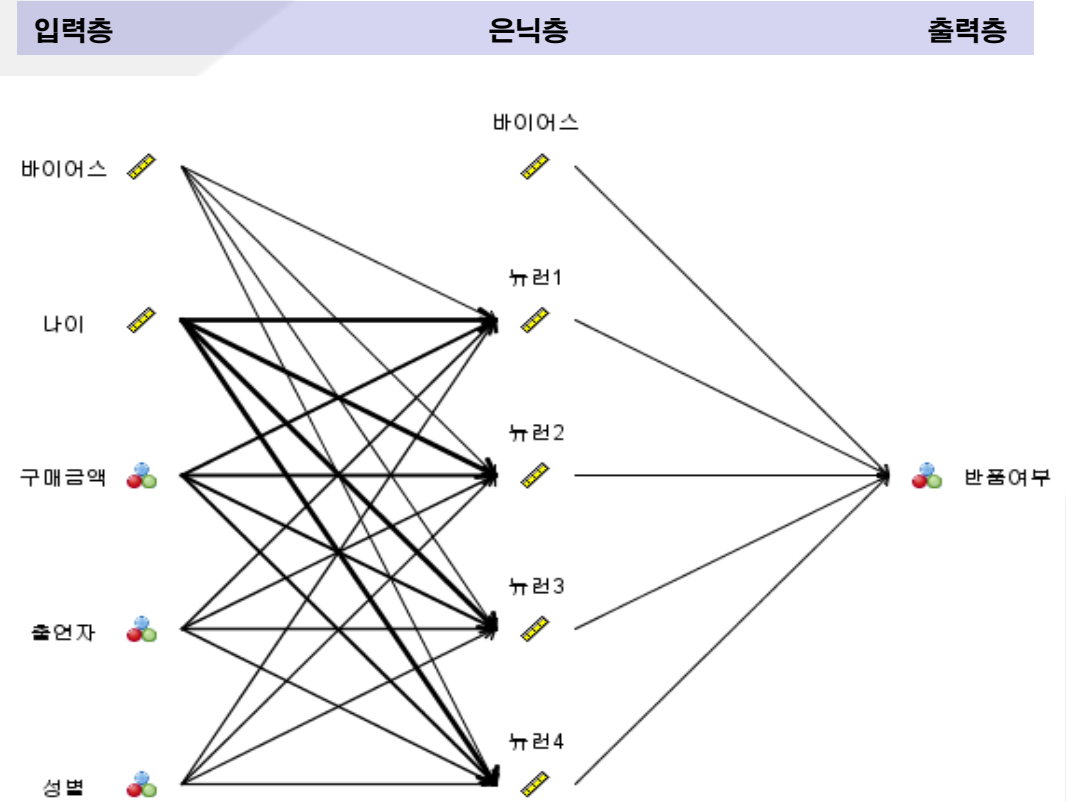
- 신경망분석 예측자 중요도
  - [신경망분석 모형] 선택 → [모형] 탭 선택 → 예측자 중요도

Neural Net 실습 (3/3)

▷ 실습 결과

- 신경망분석 네트워크
  - [신경망분석 모형] 선택 → [모형] 탭 선택 → 네트워크

▪ 신경망 분석 모형 > 모형 > 네트워크



- 1개의 은닉층 안에 4개의 뉴런으로 구성되어 있는 것을 확인할 수 있음.

- 네트워크를 시각적으로 확인할 수 있지만, 어떠한 과정으로 은닉층이 형성 되었는지에 대한 해석이 가능하지 않음.

분류 및 예측 (3) :
모형 평가 (Model Evaluation)

모형평가의 기본 개념

- 모형평가의 기준
  - 일반화의 가능성
    - 같은 모집단 내의 다른 데이터에 적용하는 경우 얼마나 안정적인 결과를 제공해 주는가?
    - 확장하여 적용가능한지 여부
  - 효율성
    - 모형이 얼마나 효과적으로 구축되었는가?
    - 얼마나 적은 입력변수로 모형을 구축했는가?
  - 예측과 분류의 정확성
    - 구축된 모형이 얼마나 예측과 분류에서 뛰어난 성능을 보이는가?
    - 아무리 안정적이고 효과적인 모형도 실제 문제에 적용했을 경우 빗나간 결과만을 양산한다면 아무런 의미가 없음
- 모형평가
  - 예측을 위해 구축된 모형이 '임의의 모형(random model)'보다 과연 우수한지, 고려된 서로 다른 모형들 중 어느 것이 가장 우수한 예측력을 보유하고 있는지 등을 비교하고 분석하는 과정
  - 성능이 좋은 모형을 찾기 위한 기준도 목표변수의 형태에 의해 다르게 고려되어야 함

모형 평가 방법 – 오분류표 (1/3)

▷ 재현율(Recall) 또는 민감도

- a/(a+b): 실제 정답인 true 중 얼마나 많은 true를 찾았는지에 대한 퍼센트

		예측 결과	
		true	false
실제	true	a	b
	false	c	d

▷ 정확도(Precision)

- a/(a+c): 모형이 true라고 판단한 것 중에서 실제 true인 것의 퍼센트

		예측 결과	
		true	false
실제	true	a	b
	false	c	d

모형 평가 방법 – 오분류표 (2/3)

▷ 특이도

- d/(c+d): 실제 정답인 false 중 얼마나 많은 false를 찾았는지에 대한 퍼센트

		예측 결과	
		true	false
실제	true	a	b
	false	c	d

▷ 정분류율(Accuracy)

- (a+d)/(a+b+c+d): 전체 결과인 a, b, c, d 중에서 실제 정답과 같은 판단을 한 퍼센트

		예측 결과	
		true	false
실제	true	a	b
	false	c	d

모형 평가 방법 - 오분류표 (3/3)

▷ 분류정확도 비교 평가의 예

- 3개의 모형 중 로지스틱 모형이 정확도 66.30%, 재현율 92.42%, 총 정확도 85.60%로 다른 모형에 비해 성과가 높다
- 로지스틱 회귀분석 > 의사결정나무분석 > 신경망분석

모형의 평가	로지스틱	의사결정 나무	인공신경망
정확도 (Precision)	66.30%	63.92%	55.36%
재현율(Recall)	92.42%	93.94%	93.94%
정분류율	85.60%	84.00%	78.40%

$L-반품여부

반품여부		0	1	합계
0	빈도	153	31	184
	행 %	83.152	16.848	100
	열 %	96.835	33.696	73.600
1	빈도	5	61	66
	행 %	7.576	92.424	100
	열 %	3.165	66.304	26.400
합계	빈도	158	92	250
	행 %	63.200	36.800	100
	열 %	100	100	100

$C-반품여부

반품여부		0	1	합계
0	빈도	149	35	184
	행 %	80.978	19.022	100
	열 %	97.386	36.082	73.600
1	빈도	4	62	66
	행 %	6.061	93.939	100
	열 %	2.614	63.918	26.400
합계	빈도	153	97	250
	행 %	61.200	38.800	100
	열 %	100	100	100

$N-반품여부

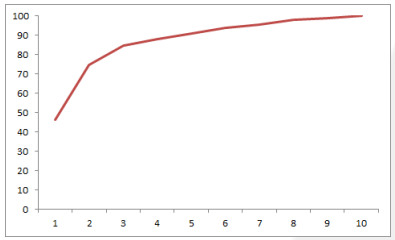
반품여부		0	1	합계
0	빈도	134	50	184
	행 %	72.826	27.174	100
	열 %	97.101	44.643	73.600
1	빈도	4	62	66
	행 %	6.061	93.939	100
	열 %	2.899	55.357	26.400
합계	빈도	138	112	250
	행 %	55.200	44.800	100
	열 %	100	100	100

모형 평가 방법 - Gains (1/2)

Gains

- 목표범주 1(true)에 속하는 개체들이 각 등급에 얼마나 분포하고 있는지를 나타냄

$$\frac{\text{해당 등급에서 목표변수의 특정 범주 빈도}}{\text{전체에서 목표변수의 특정 범주 빈도}} \times 100$$

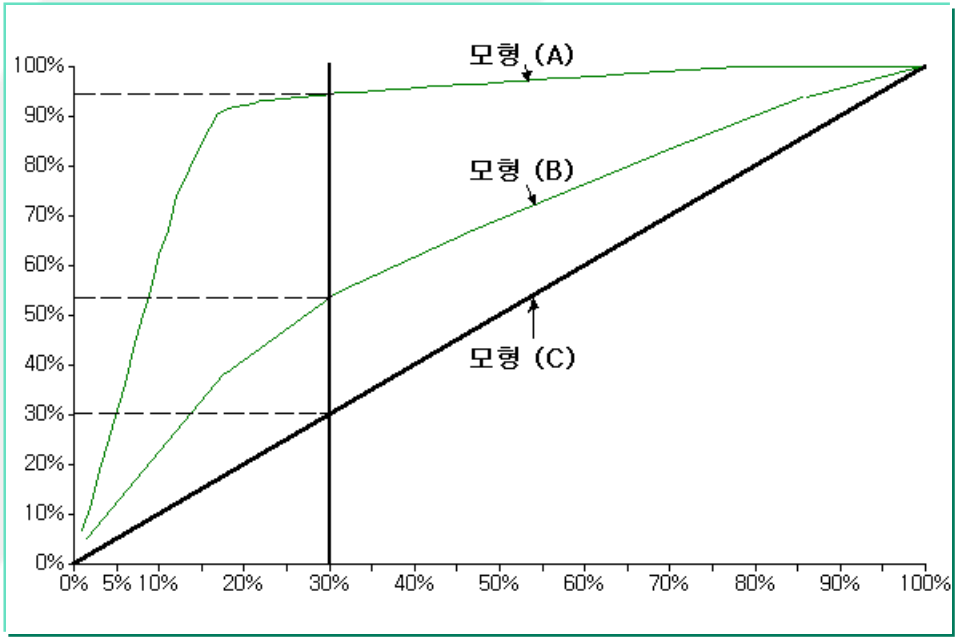


등급	비누적				누적			
	빈도			반응률	빈도			반응률
	합계	Y=1	Y=0	Gain	합계	Y=1	Y=0	Gain
1	200	174	26	**174/381=45.6**	200	174	26	**174/381=45.6**
2	200	110	90	**110/381=28.8**	400	284	116	**284/381=74.5**
3	200	38	162	**38/381= 9.9**	600	322	278	**322/381=84.5**
4	200	14	186	**14/381= 3.6**	800	336	464	**336/381=88.1**
5	200	11	189	**11/381= 2.8**	1000	347	653	**347/381=91.0**
6	200	10	190	**10/381= 2.6**	1200	357	843	**357/381=93.7**
7	200	7	193	**7/381= 1.8**	1400	364	1036	**364/381=95.5**
8	200	10	190	**10/381= 2.6**	1600	374	1226	**374/381=98.1**
9	200	3	197	**3/381= 0.7**	1800	377	1423	**377/381=98.9**
10	200	4	196	**4/381= 1.0**	2000	381	1619	**381/381=100**
전체	전체 반응률 =381/2000=19%							

모형 평가 방법 - Gains (2/2)

Gains Chart

- 해당 등급에 따라 계산 된 Gain값을 연속적으로 연결한 도표
- 차트에서 볼 수 있는 좌하에서 우상을 걸친 대각선은 모형비교의 기준선으로서, 모형성능이 나쁘면 나쁠수록 이 기준선에 가까워짐

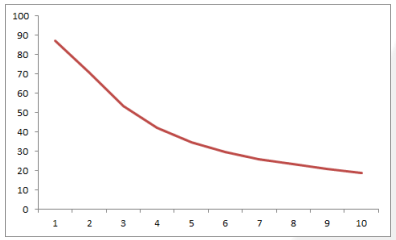


- DM 발송에 대한 반응여부라는 목표변수의 '반응했음'이라는 범주에 대한 차트라고 가정해 보자
- 서로 다른 방법으로 구축된 3개의 모형으로 동일한 수의 DM을 발송하는 경우, 얻어지는 반응률의 차이를 알 수 있음
- 전체 관찰치 중 30%를 대상으로 DM을 발송하였을 때, 모형 (A)는 90%이상의 반응을 보임
- 반면 모형 (B)는 50% 조금 넘는 반응을 보임
- 따라서 분석자는 모형 (A)를 선택하게 됨

모형 평가 방법 - Response

- **Response**
  - 각 등급에서 목표범주 1(true)의 비율을 나타냄

$$\frac{\text{해당 등급에서 목표변수의 특정 범주 빈도}}{\text{해당 등급에서 전체 빈도}} \times 100$$



등급	비누적				누적			
	빈도			반응률	빈도			반응률
	합계	Y=1	Y=0	Response	합계	Y=1	Y=0	Response
1	200	174	26	**174/200=87.0**	200	174	26	**174/200=87.0**
2	200	110	90	**110/200=55.0**	400	284	116	**284/400=71.0**
3	200	38	162	**38/200=19.0**	600	322	278	**322/600=53.6**
4	200	14	186	**14/200=7.0**	800	336	464	**336/800=42.0**
5	200	11	189	**11/200=5.5**	1000	347	653	**347/1000=34.7**
6	200	10	190	**10/200=5.0**	1200	357	843	**357/1200=29.7**
7	200	7	193	**7/200=3.5**	1400	364	1036	**364/1400=26.0**
8	200	10	190	**10/200=5.0**	1600	374	1226	**374/1600=23.3**
9	200	3	197	**3/200=1.5**	1800	377	1423	**377/1800=20.9**
10	200	4	196	**4/200=2.0**	2000	381	1619	**381/2000=19.0**

모형 평가 방법 - Lift

Lift

- 전체 반응률에 비해 각 등급에서 반응률이 얼마나 높은지를 나타냄
- 상위 등급에서의 Lift가 매우 크고 하위 등급으로 갈수록 Lift가 감소하면 이는 모형의 예측력이 적절함을 의미함 등급에 관계없이 Lift에 별 차이가 없다면 이는 모형의 예측력이 좋지 않음을 나타냄

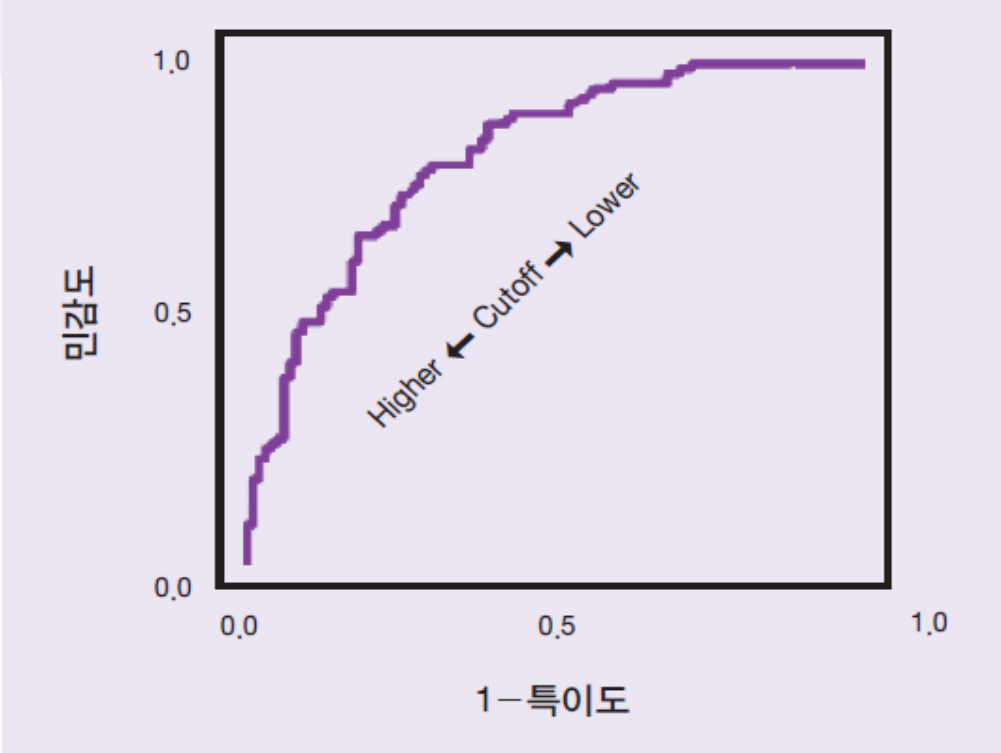
$$\frac{\text{해당 등급에서 반응률(response)}}{\text{전체 반응률}}$$

등급	비누적				누적			
	빈도			반응률	빈도			반응률
	합계	Y=1	Y=0	Lift	합계	Y=1	Y=0	Lift
1	200	174	26	**870/190=4.57**	200	174	26	**870/190=4.57**
2	200	110	90	**550/190=2.89**	400	284	116	**710/190=3.73**
3	200	38	162	**190/190=1.00**	600	322	278	**536/190=2.82**
4	200	14	186	**70/190=0.36**	800	336	464	**420/190=2.21**
5	200	11	189	**55/190=0.28**	1000	347	653	**347/190=1.82**
6	200	10	190	**50/190=0.26**	1200	357	843	**297/190=1.56**
7	200	7	193	**35/190=0.18**	1400	364	1036	**260/190=1.36**
8	200	10	190	**50/190=0.26**	1600	374	1226	**233/190=1.23**
9	200	3	197	**15/190=0.07**	1800	377	1423	**209/190=1.10**
10	200	4	196	**20/190=0.10**	2000	381	1619	**190/190=1.00**
전체	전체 반응률 =381/2000=19%							

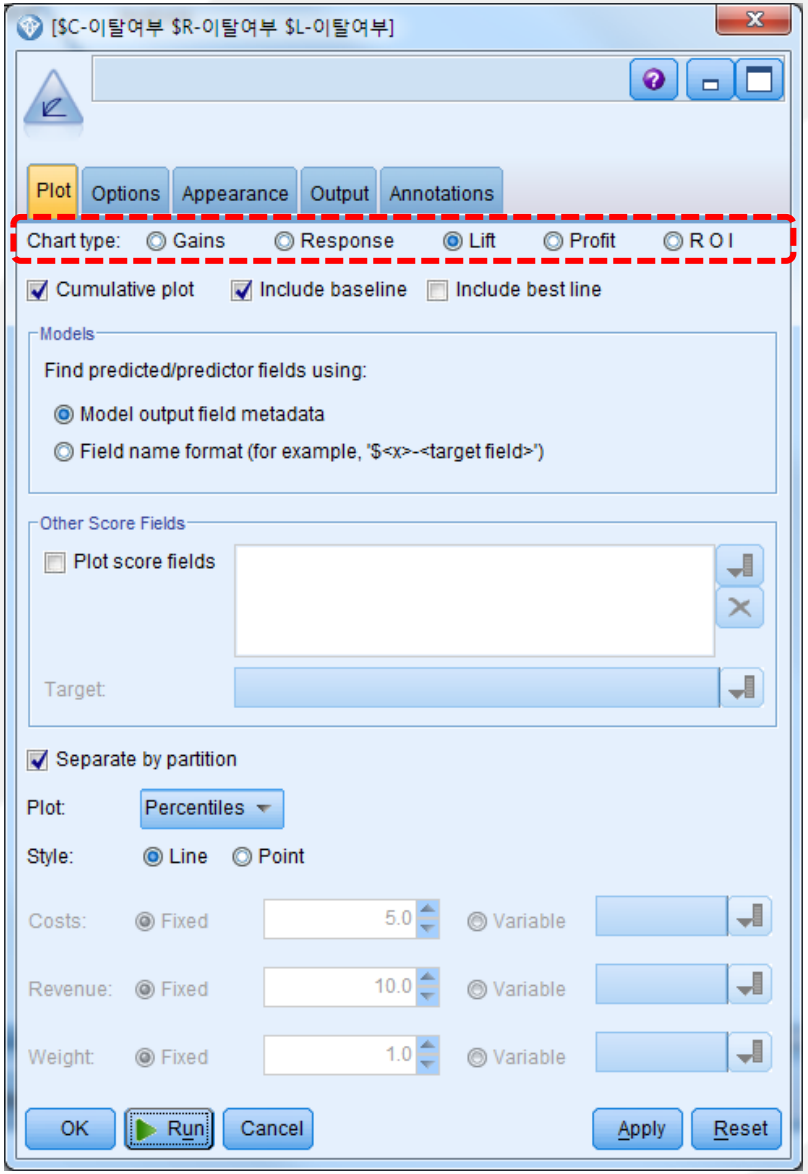
모형 평가 방법 - ROC

▷ ROC-chart

- 1-특이도(x축)와 민감도(y축)의 관계로 모형을 판단
- 모형 판단의 기준은 ROC-curve의 밑부분 면적(area under the ROC curve; AUC)이 넓을수록 좋은 모형으로 봄
  - 만약 AUC가 1이라면 완벽한 모형이라고
  - 일반적으로 덜 정확한(0.5 < AUC ≤ 0.7), 정확한(0.7 < AUC ≤ 0.9), 매우 정확한(0.9 < AUC < 1) 그리고 완벽한 모형(AUC = 1)으로 분류할 수 있음



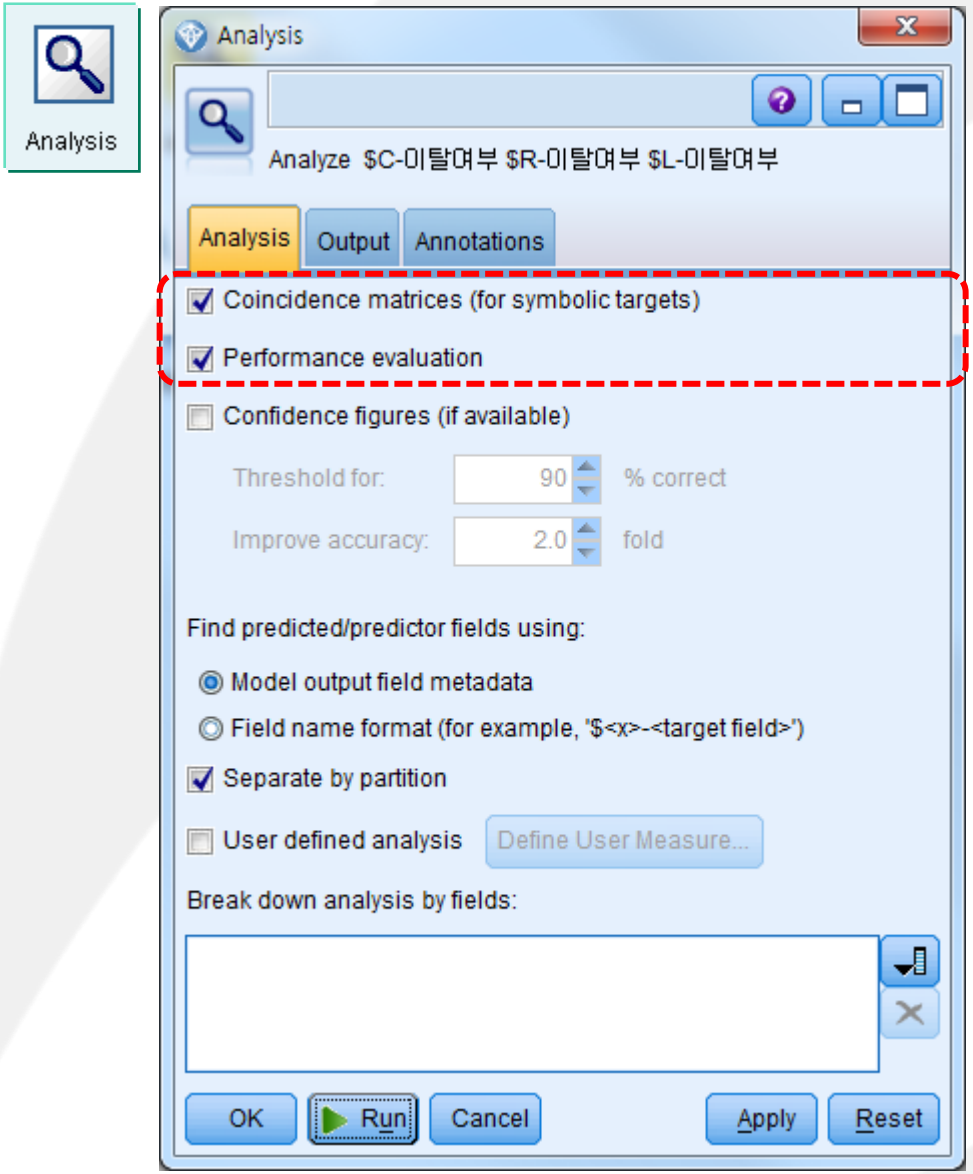
Evaluation 노드



Evaluation 노드(평가도표 노드)

- 모델링이 완료된 뒤에 모델에 대한 평가 척도를 그릴 수 있는 노드
- Evaluatin(평가) 노드의 경우에는 Generated 모델 노드(모델 실행 결과)가 만들어 진 경우에만 사용이 가능
- Chart type
  - Ganis, Response, Lift, Profit, ROI Chart
- Cumulative plot(누적도표)
  - ✓ 이를 선택하게 되면, 각 분위 그룹별로 계산이 되어진 Ganis, Response, Lift, ROI, 응답값을 계속 누적한 필드를 자체적으로 생성하고 이를 연결하여 도표를 그림
  - ✓ 평가된 내용을 분석자가 쉽게 알아보기 위해 이 옵션을 선택하는 것이 일반적임
- Include baseline
  - ✓ 그려진 평가도표를 쉽게 비교하기 위하여 일종의 참조선을 그려주는 옵션

Analysis 노드



Analysis 노드(분석 노드)

- 모델이 가진 예측력을 평가하는데 사용
- 모델 노드에서 하나 또는 그 이상 생성된 서로 다른 모델들 간의 예측 값이 실제 목표 값(Target value)을 얼마나 정확하게 예측하는지를 비교하는데 사용
- Coincidence matrices(for symbolic targets)(일치도 교차표(문자형 목표 필드의 경우)
  - ✓ 문자형 목표 필드에 대하여 각각의 생성된 예측 필드와 목표 필드 사이의 대응 패턴을 보여줌
  - ✓ 동일한 목표 필드에 대해 서로 다른 예측 모델이 존재하면 분석 노드는 모델 간 예측 값이 일치한 케이스 예측 값이 모두 적중한 케이스 및 전체 합계 등을 보여줌
- Performance evaluation(성능평가)
  - ✓ 문자형 출력 필드를 갖는 모델에 대한 성능 평가 통계량을 제공함

분류 및 예측 (4) :
Automated Modeling & Ensemble

Automated Modeling

Supervised Modeling(통제학습/관리학습)을 실시할 경우, 목표변수가 범주형인지 연속형인지에 따라 분석 방법을 달리 하면서 최적의 모델은 선정하게 됨

Automated Modeling 방법을 통하여 목표변수가 범주형일 경우, Auto Classifier 노드를 사용하고 연속형일 경우, Auto Numeric 노드를 사용하여 최적의 모형을 선정함

Auto Classifier 지원 가능 모형

- Neural Net
- C5.0
- C&R Tree
- QUEST
- CHAID
- Logistic
- Decision List
- Bayes Net
- Discriminant
- SVM

신경망, 의사결정나무(C5.0, CART, QUEST, CHAID), 로지스틱 회귀분석, 판별분석, Bayes Net, 의사결정 리스트 모형의 이분형 분류 결과를 비교할 수 있다. 비교를 원하는 모형을 선택할 수 있고, 사용하고자 하는 알고리즘과 각 알고리즘을 위한 상세한 옵션을 선택할 수 있습니다. 모델들은 Overall accuracy, Area under the curve, Profit, Lift, Number of variables를 기초해 비교된다.

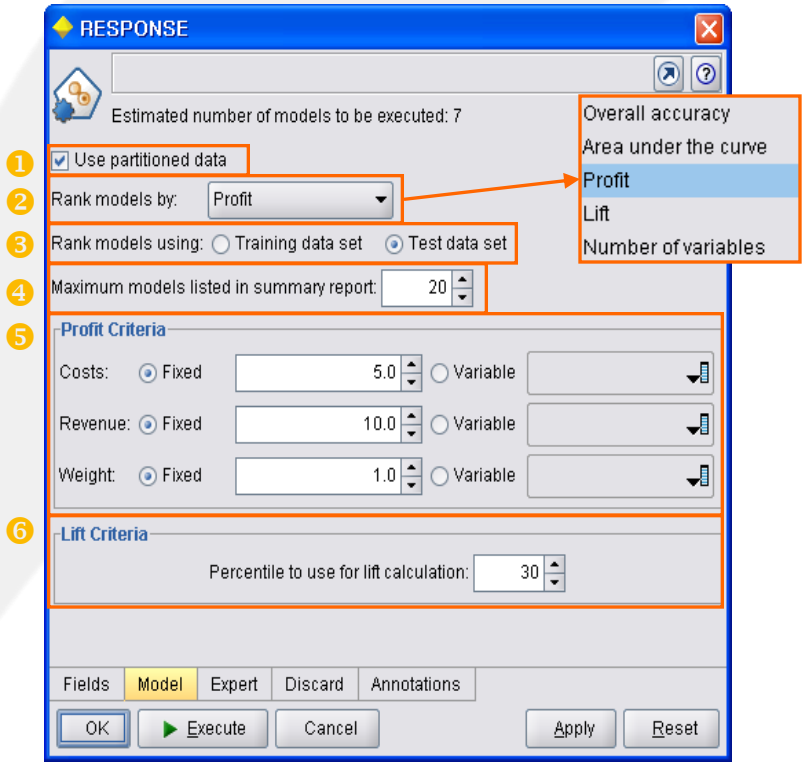
Auto Numeric 지원 가능 모형

- Neural Net
- C&R Tree
- CHAID
- Regression
- Generalize...
- SVM

신경망, 의사결정나무(CART, CHAID), 회귀분석, 일반화선형모형, SVM 모형의 예측 결과들을 비교할 수 있다. 이 노드는 사용할 알고리즘을 선택하고 각 모델링 과정에서 옵션들의 다양한 조합을 실험해보는 Binary Classifier노드와 같은 방법을 사용한다. 모델들은 상관관계, 상대오차, 사용된 변수의 수 등에 기초해 비교된다.

Auto Classifier Node (1/4)

- Binary Classifier Node는 다양한 이분형 분류 모형을 생성하고, 그 결과를 비교하여 주는 Node입니다.
- 원하는 모형을 선택할 수 있으며, 각 모형 별로 모형 모수 및 세부 옵션을 지정할 수 있습니다.



Model Tab

❶ 데이터들을 현재 partition 필드에 기준해서 training, testing, validation에 대한 샘플 또는 하위세트로 나누어줌

❷ 모형을 비교하는 기준을 정의.

- Overall accuracy : 정확도
- Area under the curve : ROC 곡선 아래 면적
- Profit : 이익의 누적합
- Lift : 전체 샘플에 대한 누적 백분위의 적중률
- Number of variables : 분석에 사용한 변수의 수

❸ 데이터에 파티션 처리가 되어 있을때, 순위를 결정하는데 트레이닝 또는 테스트 데이터셋을 사용할지 선택.

❹ 노드에서 작성되는 요약 보고서에 나타나는 최대 모형 개수.

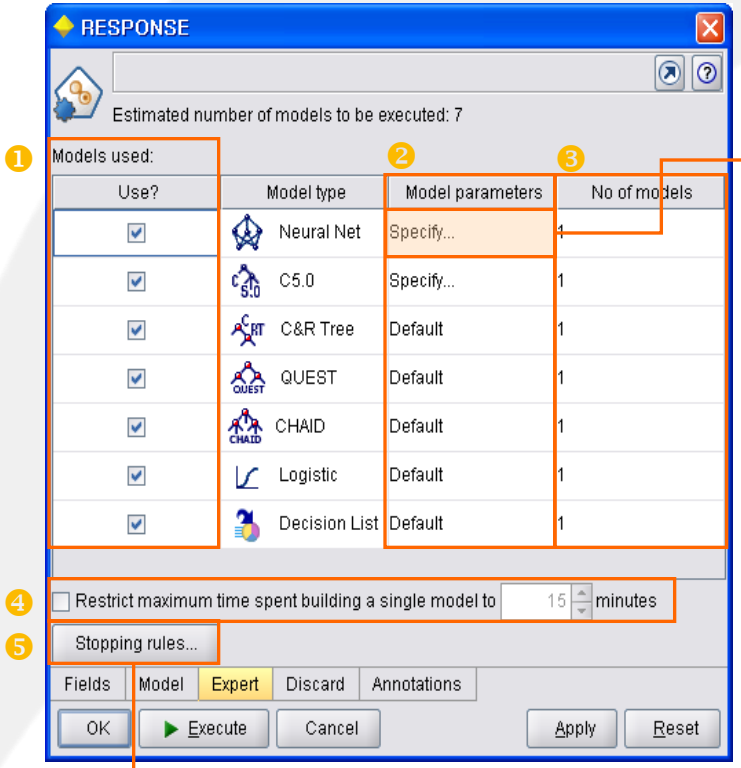
❺ 이익을 산출하기 위한 값을 지정.

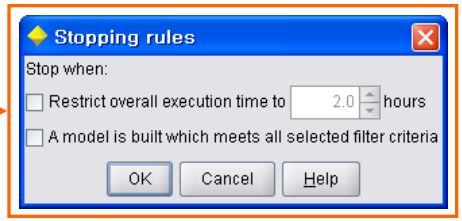
- Costs : 레코드의 비용을 정의. 비용 변수가 따로 있으면 fixed 대신 variable을 선택한 다음, 해당 변수를 지정.
- Revenue : 레코드의 수익을 정의. 수익 변수가 따로 있으면 fixed 대신 variable을 선택한 다음, 해당 변수를 지정.
- Weight : 데이터가 가중케이스로 표현되어 있으면 가중케이스 변수를 지정.

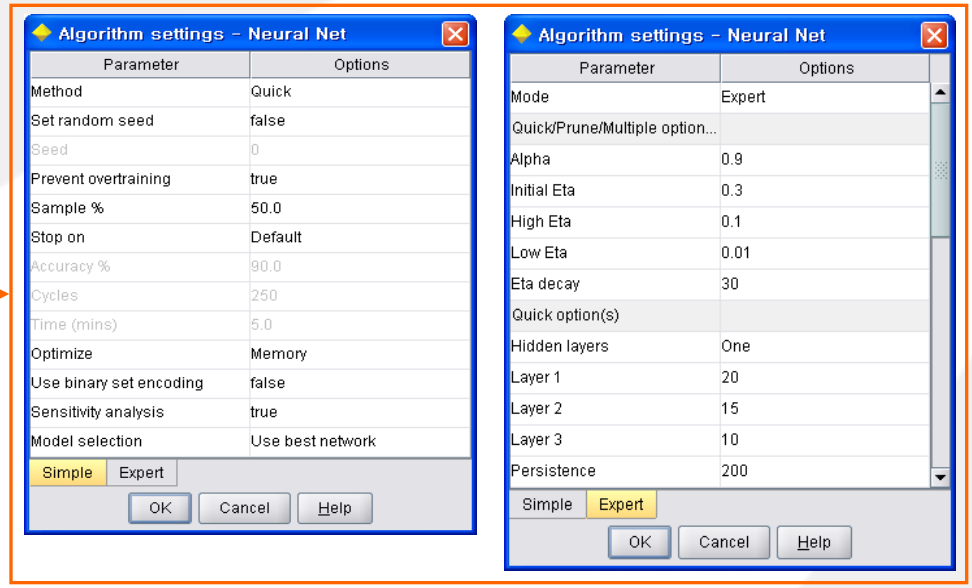
❻ 리프트 계산을 위한 백분율 지정

Auto Classifier Node (2/4)

- Expert tab에서는 사용할 알고리즘을 선택하고 정지 규칙을 지정할 수 있습니다.
- 선택된 각 모형의 모수를 사용자가 지정할 수 있습니다.







❶ 비교하고자 하는 모형을 왼쪽의 체크박스에서 선택.

❷ 각 모형에서 디폴트 세팅을 그대로 적용할 수도 있고, 'Specify...'를 선택하여 옵션을 선택할 수 있음. 특정 옵션은 각 모형 노드와 유사하며 여러가지 모수 조합을 하나의 노드에서 설정하여 비교 가능함.

❸ Model parameters 지정에 따라 작성되는 모형의 수를 출력.

Expert Tab

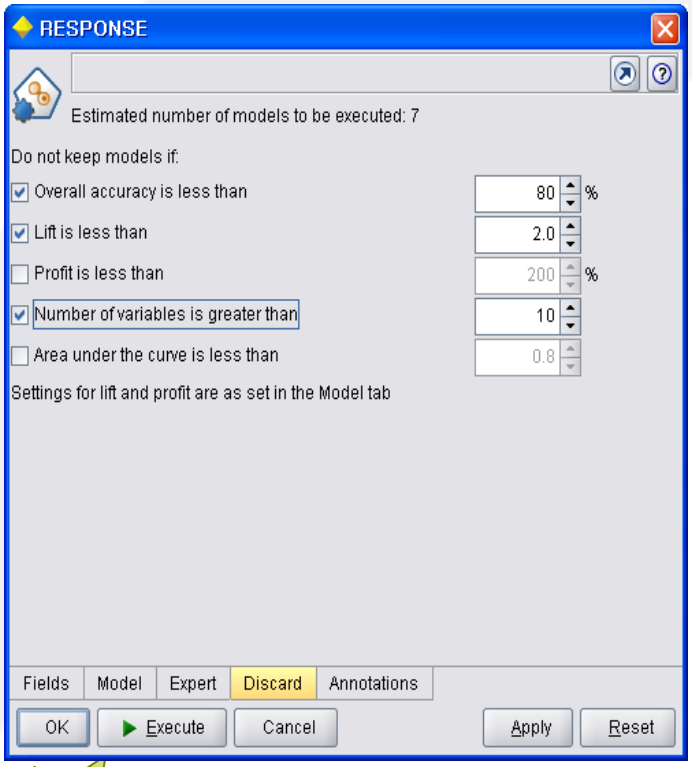
❹ 하나의 모형을 작성하는데 걸리는 시간의 최대값 지정.

❺ 전체 노드를 실행하는 것에 대해 Binary Classifier 노드의 정지 규칙을 정의.

- Restrict overall execution time to : 지정된 시간 이후 정지.
- A model is built that meets all selected filter criteria : Discard tab에서 지정된 기준을 모두 수행한 뒤 정지.

Auto Classifier Node (3/4)

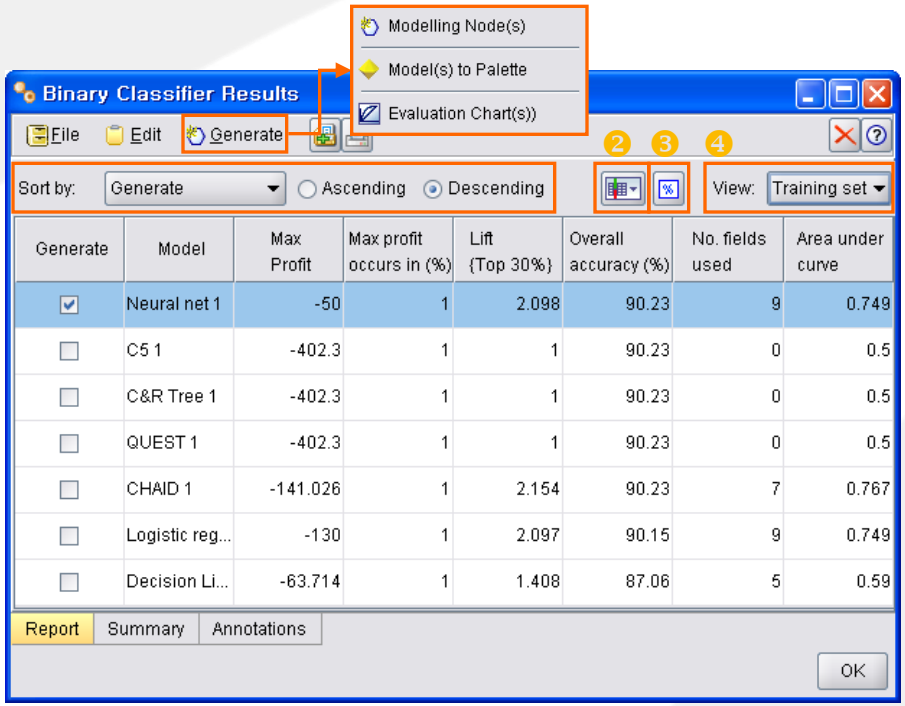
- Discard tab에서는 제외 규칙을 정의하고, 규칙에 맞지 않는 모형은 자동으로 제외하고 모형을 비교합니다.



Discard Tab

✓ 정확도, 리프트, 이익의 최소값과 사용된 변수의 최대값, 곡선 아래 면적의 최소값을 지정.

실행



Generate	Model	Max Profit	Max profit occurs in (%)	Lift {Top 30%}	Overall accuracy (%)	No. fields used	Area under curve
☑	Neural net 1	-50	1	2.098	90.23	9	0.749
☐	C5 1	-402.3	1	1	90.23	0	0.5
☐	C&R Tree 1	-402.3	1	1	90.23	0	0.5
☐	QUEST 1	-402.3	1	1	90.23	0	0.5
☐	CHAID 1	-141.026	1	2.154	90.23	7	0.767
☐	Logistic reg...	-130	1	2.097	90.15	9	0.749
☐	Decision Li...	-63.714	1	1.408	87.06	5	0.59

Results Browser

✓ Binary Classifier 결과 창에서는 모형 생성 시간, 이익, 리프트, 정확도 등의 모형 실행 결과를 요약적으로 보여줍니다.

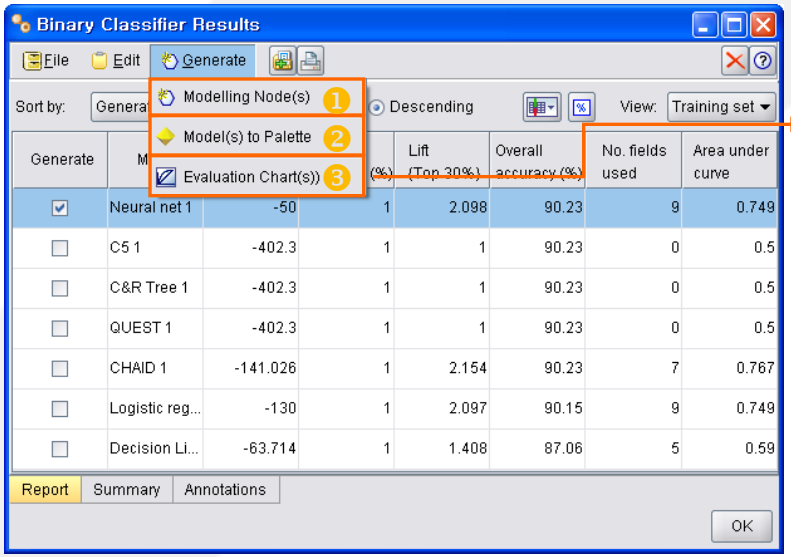
❶ 결과 창에서 지정한 열에 대해 모형을 정렬할 수 있습니다.

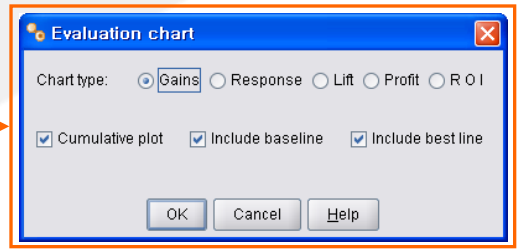
❷ 특정 열만 선택적으로 출력할 수 있습니다.

❸ 리프트 계산을 위한 백분율 값을 수정할 수 있습니다.

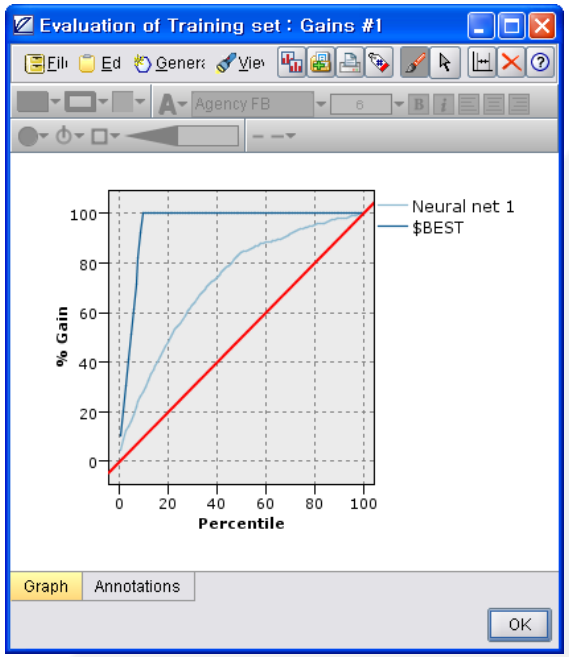
❹ 데이터들을 현재 partition 필드에 기준해서 training, testing, validation에 대한 샘플 또는 하위세트로 나누어줌

Auto Classifier Node (4/4)





실행



Results Browser

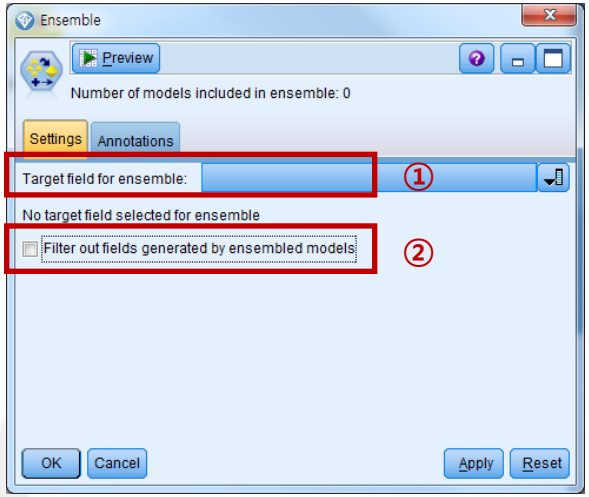
✓ Generating Nodes and Models

❶ 선택한 모형의 노드가 팔레트에 생성됩니다.

❷ 모형 생성 결과가 팔레트에 생성됩니다.

❸ 모형 평가 도표를 작성합니다.

Ensemble 노드



- Ensemble 노드는 둘 이상의 모델 광석(nugget)을 결합해 각각의 모델을 이용해 얻을 수 있는 것보다 더 정확한 예측값을 얻기 위해 사용된다.
  - 여러 개의 모델들로부터의 예측값을 결합함으로써, 각각의 모델들의 제한사항들을 피할 수 있고, 전반적으로 더 높은 정확도를 얻을 수 있다.
  - 이 방법으로 결합된 모델들은 일반적으로 각각의 모델들 중 가장 좋은 모델 이상의 성과를 낼 수 있다.
- 자동화 모델링 노드를 사용할 때 자동으로 Ensemble이 적용된다.

① Ensemble 노드에서 목표변수 설정
② "Filter out fields generated by ensembled models"을 선택해제한다.
  - 이를 통해 결합된 ensemble 스코어와 각각의 모델의 스코어를 비교할 수 있음.

Example 1. 회사는 각 고객에게 알맞은 제안을 함으로써 더욱 이익을 낼 수 있는 결과를 얻기를 원한다. Binary Classifier 노드를 이용하여 고객이 특정 제안을 받아들일지 예측하는 여러 개의 모델을 만들고 Ensemble 노드를 이용해 이 모델들로부터 하나의 결합된 스코어를 생성한다.
Example 2. 지방자치단체에서 더욱 정확하게 부동산 세를 추정하고, 모든 자산을 조사해 볼 필요없이 특정한 자산에 대한 값만을 조정하기를 원한다면, Numeric Predictor 노드를 이용하여 분석가는 건물 유형, 주거 환경, 규모, 기타 다른 요소들을 이용하여 자산의 가치를 예측할 수 잇는 수많은 모델을 생성하고 비교해 볼 수 있으며, Ensemble 노드를 이용해 이 스코어들을 결합할 수 있다.
Ensemble 노드를 사용한 후에 Analysis 노드나 Evaluation 노드를 이용하여 하나로 결합된 결과의 정확도를 입력된 모델들과 비교해 볼 수 있다. 이를 위해, Ensemble 노드의 Setting 탭에서 "Filter out fields generated by ensembled models"을 선택해제 해 줘야 한다.
Output Fields. 각 Ensemble 노드는 결합된 스코어를 보여주는 변수를 생성한다. 이 변수들의 이름은 목표변수에 기초하며 변수의 유형(flag, nominal, continuous)에 따라 $XF-, $XS-, $XR-의 접두어가 붙는다. 예를 들어 "response"란 이름의 flag 유형의 목표변수가 있다면 생성되는 변수의 이름은 "$XF-response"이 될 것이다.